446 Section 3.000001

TA: Yufei Zhang

Plans for today!

1. This
2. Matrix Vector Proof
3. Vector Calculus
4. Approximations
5. Problem 1.2

Today's section is going to be *super* math heavy…

It's okay if not everything makes sense right away!

Our goal is to develop *intuition* for the math :)

Plans for today!

1. This
2. Matrix Vector Proof
3. Vector Calculus
4. Approximations
5. Problem 1.2

Reminders

- **HW1** due **Wed, Jan 28**

Aside (quick matrix proof)

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad X = \begin{bmatrix} - & x_1^\top & - \\ - & x_2^\top & - \\ & \vdots & \\ - & x_m^\top & - \end{bmatrix}, \quad w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

$$Xw - Y = \begin{bmatrix} x_1^\top w - y_1 \\ x_2^\top w - y_2 \\ \vdots \\ x_m^\top w - y_m \end{bmatrix}$$

$$\|Xw - Y\|_2^2 = \sum_{i=1}^{m} (x_i^\top w - y_i)^2$$

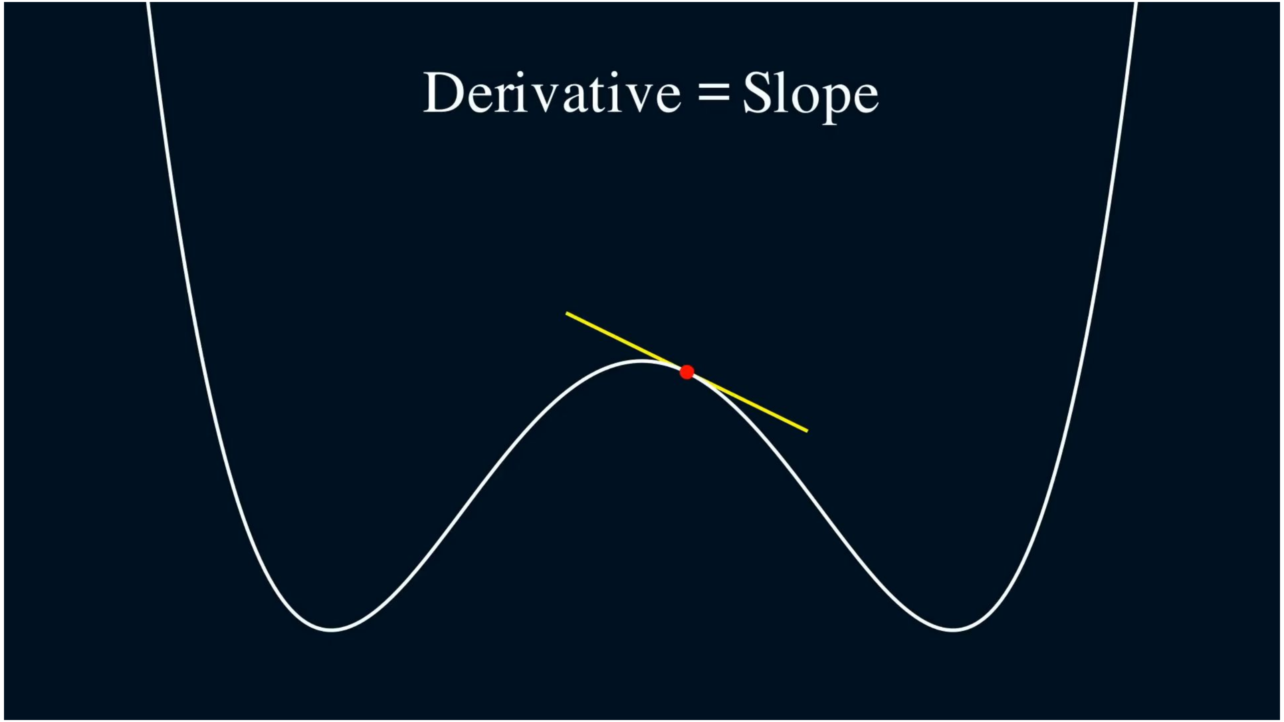
$(a-b)^2 = (b-a)^2$. Therefore:

$$(x_i^\top w - y_i)^2 = (y_i - x_i^\top w)^2$$

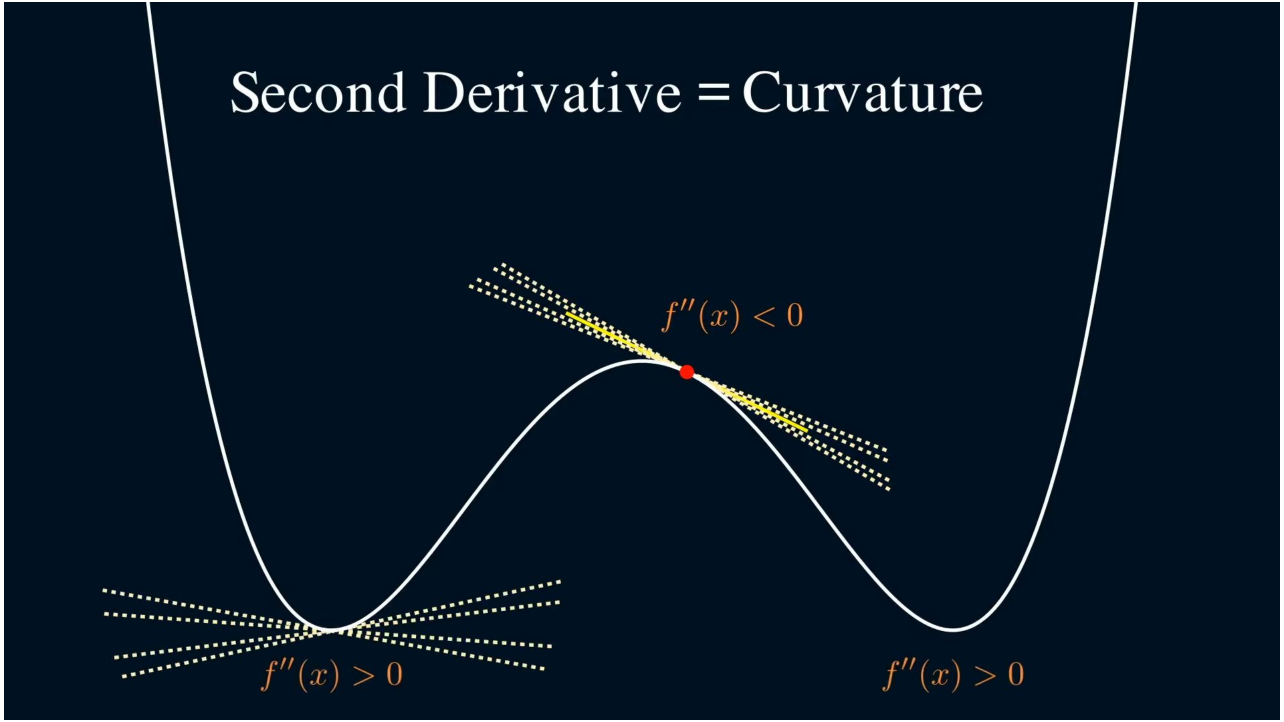
$$\|Xw - Y\|_2^2 = \sum_{i=1}^{m} (y_i - x_i^\top w)^2$$

Derivative

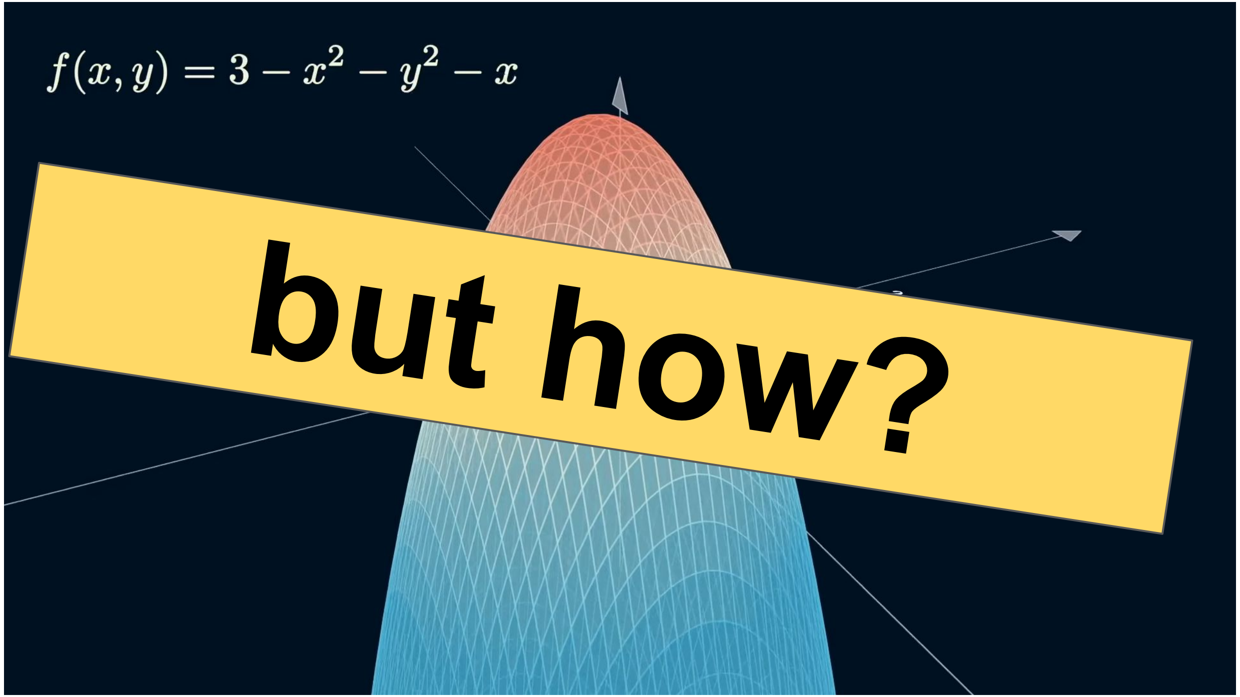
What is a derivative?



What is a second derivative?

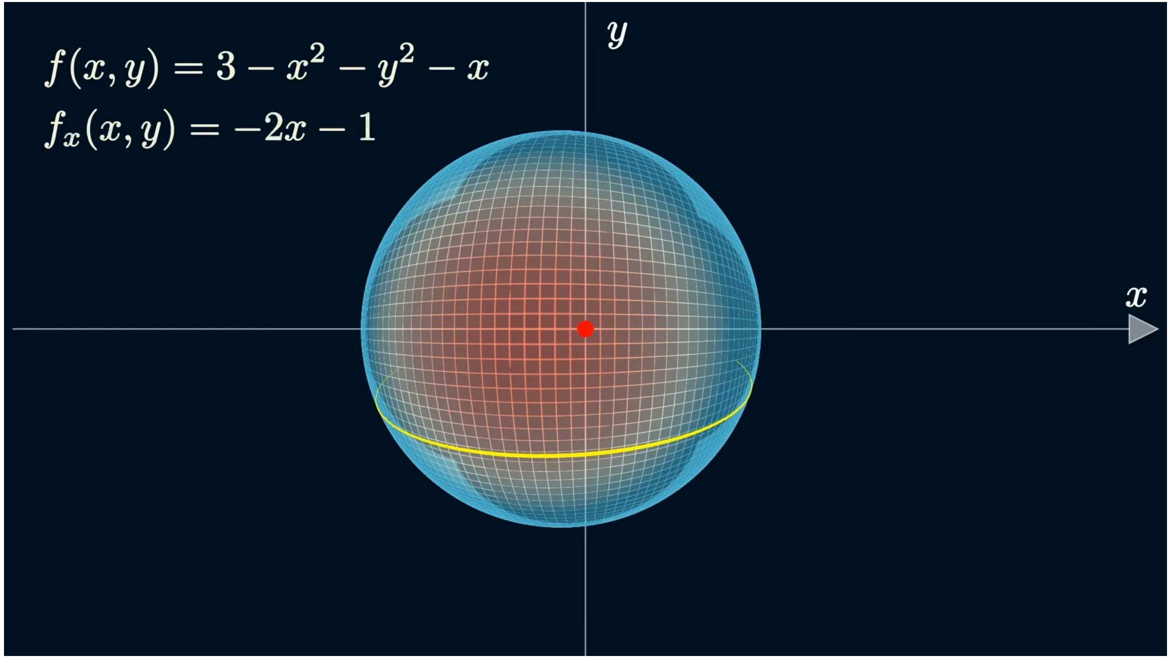


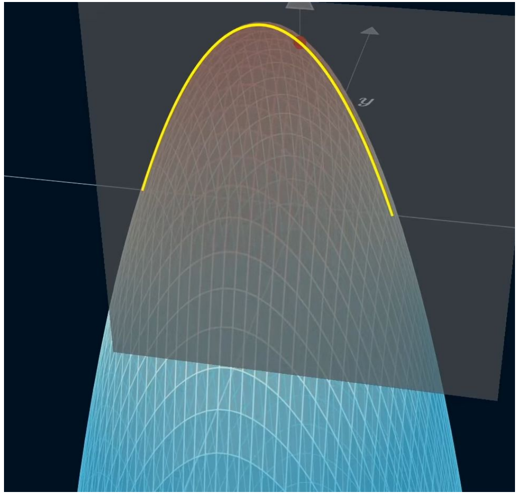
Find the derivatives along different directions in this graph



Calculate $f_x(x, y)$

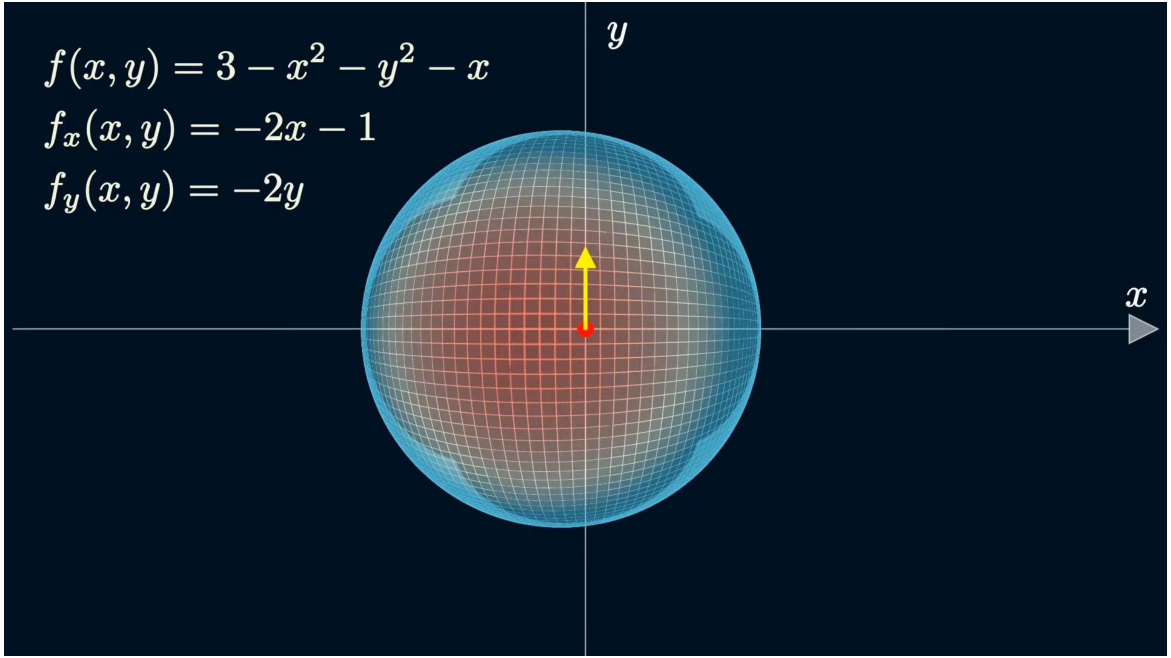
Treat **y** as a constant and take the partial derivative wrt to **x**





Calculate $f_y(x, y)$

Treat **x** as a constant and take the partial derivative wrt to **y**



Gradient

What is the gradient?

Let **f** be a scalar-valued multivariable function **f(x, y, …)**

The **gradient of f** is the collection of **f's partial derivatives** in a vector:

Scalar-valued multivariable function

$$\nabla f(x_0, y_0, \dots) = \begin{bmatrix} \frac{\partial f}{\partial x}(x_0, y_0, \dots) \\ \frac{\partial f}{\partial y}(x_0, y_0, \dots) \\ \vdots \end{bmatrix}$$

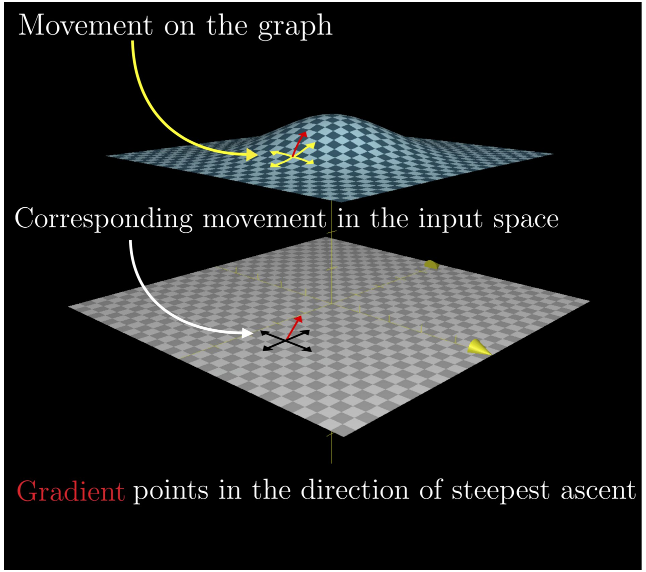
∇f takes the same type of inputs as f

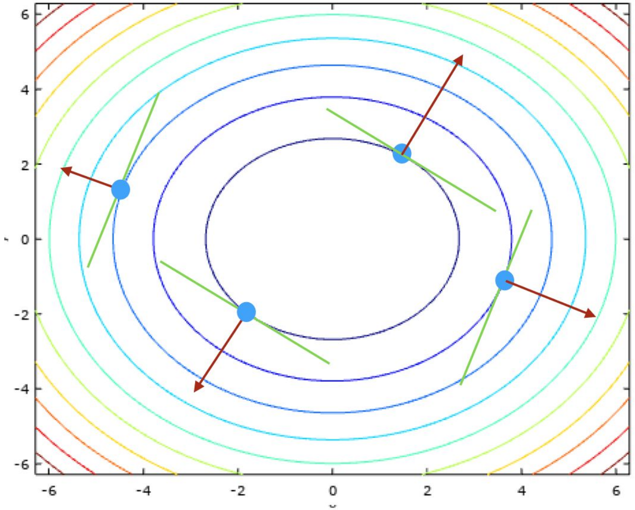
Notation for gradient, called "nabla".

∇f outputs a vector with all possible partial derivatives of f.

Gradient in multiple dimensions

The *gradient vector* of a function of several variables at any point **denotes the direction of maximum rate of change**





Gradient vectors at various points shown with red arrows
Tangent to the contour is in green

Calculating the gradient

Input:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

Function:

$$f(x) : \mathbb{R}^n \to \mathbb{R}$$

Calculating the gradient

Input:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix}$$

Function:

$$f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$$

Take the partial derivative n times:

$$\Delta_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix}$$

Vocab

Name	Symbol	Example
Derivative	$\frac{d}{dx}$	$\frac{d}{dx}(x^2) = 2x$
Partial derivative	$\frac{\partial}{\partial x}$	$\frac{\partial}{\partial x}(x^2 - xy) = 2x - y$
Gradient	∇	$\nabla(x^2 - xy) = \begin{bmatrix} 2x - y \\ -x \end{bmatrix}$

How to calculate the gradient

Let's take an example. I have a function defined as $f(x, y) = 5x^2 + 3xy + 3y^3$. First, we need to find the partial derivatives with respect to the variables x and y as follows:

$$\frac{\partial f}{\partial x} = 10x + 3y$$
$$\frac{\partial f}{\partial y} = 3x + 9y^2$$

This gives us a gradient:

$$\nabla f = \begin{bmatrix} 10x + 3y \\ 3x + 9y^2 \end{bmatrix}$$

Jacobian

From Gradient → Jacobian

$$f : \mathbb{R}^u \to \mathbb{R} \qquad \mathbf{J} = \frac{df(x)}{dx} = \left[\frac{\partial f(x)}{\partial x_1} \dots \frac{\partial f(x)}{\partial x_u} \right]$$

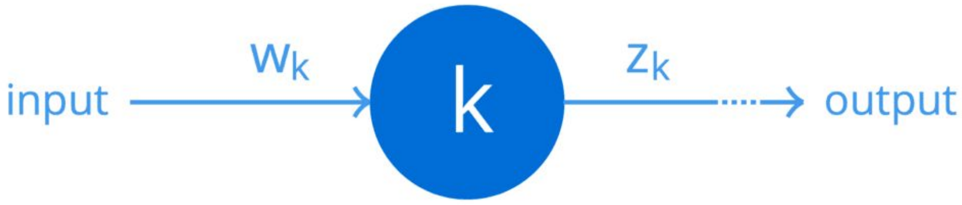
these are the same ***values*** as the gradient!

if we generalize this function to more dimensions…

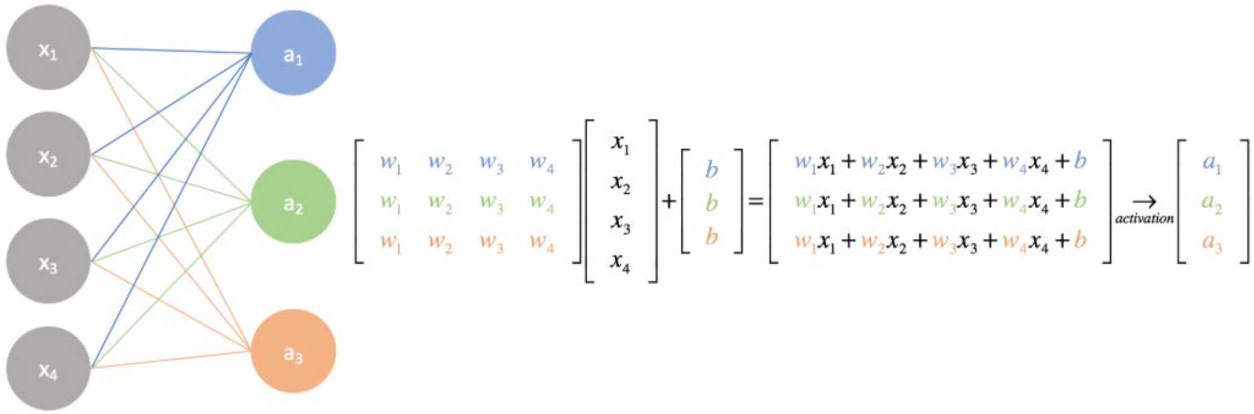
$$\mathbf{f} : \mathbb{R}^u \to \mathbb{R}^v \qquad \mathbf{J} = \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \left[\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} \dots \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_u} \right] = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_u} \\ \vdots & & \vdots \\ \frac{\partial f_v(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_v(\mathbf{x})}{\partial x_u} \end{bmatrix}$$

From Gradient → Jacobian

$f : \mathbb{R}^u \to \mathbb{R}$



$\mathbf{f} : \mathbb{R}^u \to \mathbb{R}^v$



Interpreting the Jacobian

How do we interpret the jacobian matrix?

$$\mathbf{J} = \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \left[\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} \cdots \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_u}\right] = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_u} \\ \vdots & & \vdots \\ \frac{\partial f_v(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_v(\mathbf{x})}{\partial x_u} \end{bmatrix}$$

This matrix gives tells us how the outputs will change when we vary the value of $\mathbf{x}_i$

For example, if we increase x_1, how is g(x) affected?

Hessian

What is the Hessian?

The hessian matrix of a multivariable function **f** organizes all **second partial derivatives** into a matrix

$$\mathbf{H}f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} & \cdots \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} & \cdots \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

What is the Hessian?

The matrix can be evaluated at some **point** $(x_0, y_0, \ldots)$ in the domain of **f**

$$\mathbf{H}f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} & \cdots \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} & \cdots \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$\mathbf{H}f(x_0, y_0, \ldots) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2}(x_0, y_0, \ldots) & \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0, \ldots) & \cdots \\ \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0, \ldots) & \frac{\partial^2 f}{\partial y^2}(x_0, y_0, \ldots) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Developing Intuition

We started with a function f that takes n inputs (a vector) → gives you 1 output

- This could be the **loss value**

We take the derivative (gradient) of this scalar function → get a vector of size n

- This vector tells you the **slope in every direction**

$$\begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

Developing Intuition

We started with a function f that takes n inputs (a vector) → gives you 1 output

- This could be the **loss value**

We take the derivative (gradient) of this scalar function → get a vector of size n

- This vector tells you the **slope in every direction**

$$\begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

What happens if we differentiate the gradient itself?

- We can't take the *gradient* of a vector → vector function
- We have to use the jacobian!
- Therefore, the Hessian Matrix is the Jacobian Matrix of the Gradient Vector.

Problems 1.2 a-b

Solve them! Ask for help if you are stuck. Look at section 1.1 for help remembering how these gradients, Jacobians, and Hessians compute.

(a) Let $f(x_1, x_2) = x_1^2 + e^{x_1 x_2} + 2\log(x_2)$. What are the gradient and the Hessian of f?

(b) Note that $\nabla_x f : \mathbb{R}^n \to \mathbb{R}^n$. What is the Jacobian of $\nabla_x f$?

Answers

(a) Let $f(x_1, x_2) = x_1^2 + e^{x_1 x_2} + 2\log(x_2)$. What are the gradient and the Hessian of f?

Solution:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 + x_2 e^{x_1 x_2} \\ x_1 e^{x_1 x_2} + \frac{2}{x_2} \end{bmatrix} \text{ and } \nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 + x_2^2 e^{x_1 x_2} & e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} \\ e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} & x_1^2 e^{x_1 x_2} - \frac{2}{x_2^2} \end{bmatrix}$$

(b) Note that $\nabla_x f : \mathbb{R}^n \to \mathbb{R}^n$. What is the Jacobian of $\nabla_x f$?

Equivalent

Solution:

$$\nabla_x(\nabla_x f)(x) = \begin{bmatrix} \frac{\partial(\nabla_x f)_1(x)}{\partial x_1} & \frac{\partial(\nabla_x f)_1(x)}{\partial x_2} \\ \frac{\partial(\nabla_x f)_2(x)}{\partial x_1} & \frac{\partial(\nabla_x f)_2(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2 + x_2^2 e^{x_1 x_2} & e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} \\ e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} & x_1^2 e^{x_1 x_2} - \frac{2}{x_2^2} \end{bmatrix} = \nabla_x^2 f(x)$$

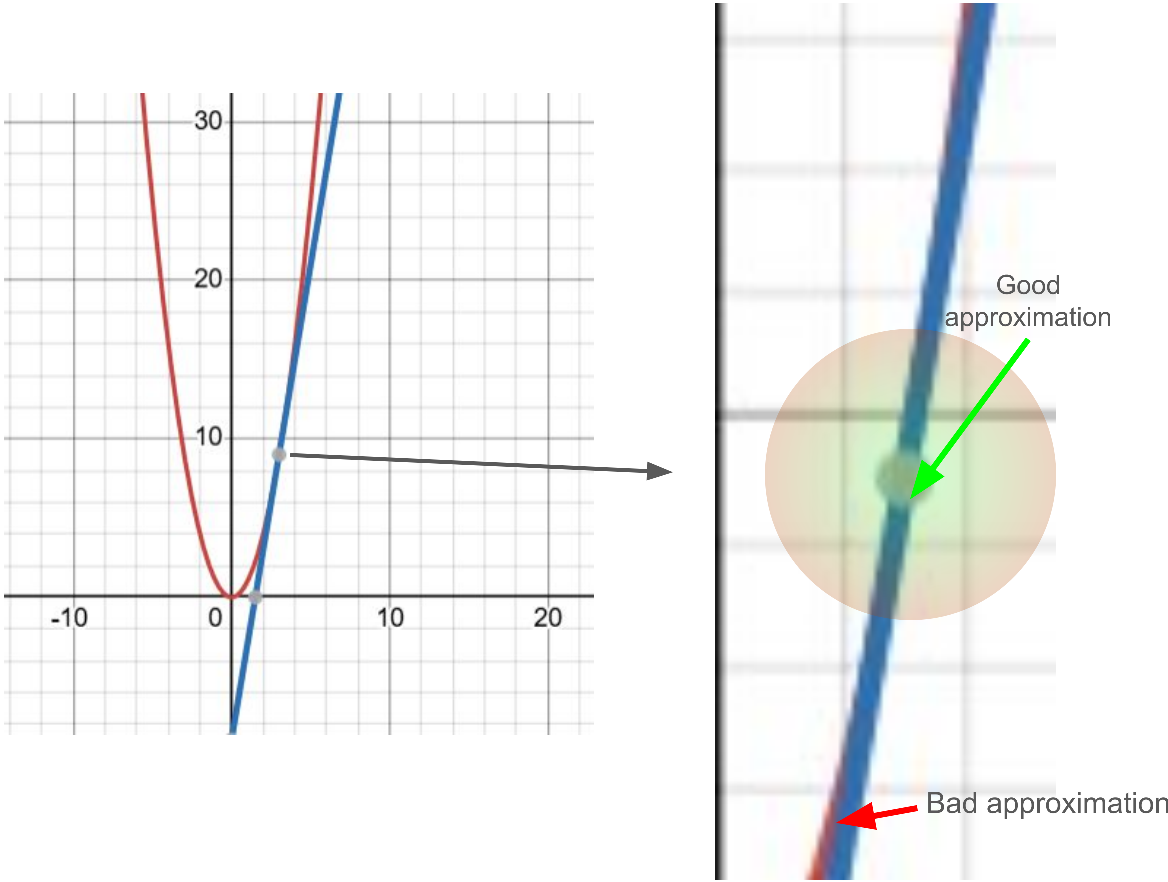
Approximations

Linear Approximation

The derivative of **f(x)** at some **(x,y)** can be used to linearly approximate **f(x ± ε)**

Where ε is very tiny!

This extends to multivariate functions… proof in your notes



Linear Approximation

For a "many-to-one" function, the gradient gives us a vector we can use to linearly approximate a small area around some **x**

$$f : \mathbb{R}^n \rightarrow \mathbb{R}$$

Let $\epsilon = [\epsilon_1, \ldots, \epsilon_n]^T$ and $x = [x_1, \ldots, x_n]^T$

$$f(x + \epsilon) \approx f(x) + \nabla_x f(x)^T \epsilon$$

What about a "many-to-many" function?

$$g : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

Problem 1.2 c

Remember that the Jacobian is just the gradient of a "many-to-many" function.

Also remember: *For a "many-to-one" function, the gradient gives us a vector we can use to linearly approximate a small area around some* ***x***

(c) The gradient $\nabla_x f(x)$ offers the best linear approximation of f around the point x. What does the Jacobian of a function $g : \mathbb{R}^n \to \mathbb{R}^m$ offer?

Answer

(c) The gradient $\nabla_x f(x)$ offers the best linear approximation of f around the point x. What does the Jacobian of a function $g : \mathbb{R}^n \to \mathbb{R}^m$ offer?

Solution:

The Jacobian also offers the best linear approximation of g around a point x, but now it approximates a vector, instead of a scalar,

$$g(x + \epsilon) \approx g(x) + \nabla_x g(x)\epsilon$$

where $\nabla_x g(x)\epsilon$ is a matrix multiplication instead of a dot product.

Problem 1.2 d

(d) If we use the gradient and the Hessian of $f : \mathbb{R}^n \to \mathbb{R}$, what type of an approximation for the function f around a point x can we create.

Remember Taylor expansion?

↳ To approximate a function around a point a

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 \dots$$

↑ Exact at a, close around a

Better and better approximations

Remember Taylor expansion?

↳ To approximate a function around a point a

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 \dots$$

↑ Exact at a, close around a

Better and better approximations

Set $a = x$, we want to estimate $x + \epsilon$

$$f(x+\epsilon) \approx f(x) + \frac{f'(x)}{1!}(x+\epsilon-x) + \frac{f''(x)}{2!}(x+\epsilon-x)^2 + \dots$$

⇓

$$f(x+\epsilon) \approx f(x) + f'(x)\epsilon + \frac{1}{2}f''(x)\epsilon^2 + \dots$$

Generalizing to vectors: $f: \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{R}^n$, $\epsilon \in \mathbb{R}^n$

$$f(x+\epsilon) \approx f(x) + (\nabla_x f(x))^T \epsilon$$

Gradient = first order derivative of $f(x)$

So what is the second order derivative?

Second order derivative = $\nabla_x (\nabla_x f(x))$ = **Hessian**

↳ Gives us a Quadratic Approximation

2nd order Taylor expansion around **x** generalized to vectors

$$f(x+\epsilon) \approx f(x) + (\nabla_x f(x))^T \epsilon + \frac{1}{2} \epsilon^T (\nabla_x^2 f(x))^T \epsilon$$

Answer!

Problem 1.2 g (IMPORTANT!)

(g) Draw the gradient on the picture. Describe what happens to the values of the approximation of f if we move from x in directions d_1, d_2, d_3 for which $\nabla_x f(x)^T d_1 > 0, \nabla_x f(x)^T d_2 < 0, \nabla_x f(x)^T d_3 = 0$? Can the same conclusions be drawn about the function of f?

$$(\nabla_x f(x))^T d_1 > 0$$

↳ Direction d_1 points generally toward the gradient

$$(\nabla_x f(x))^T d_2 < 0$$

↳ Direction d_2 points generally away from the gradient

$$(\nabla_x f(x))^T d_3 = 0$$

↳ Direction d_3 points orthogonal to the gradient

Answer

(g) Draw the gradient on the picture. Describe what happens to the values of the approximation of f if we move from x in directions d_1, d_2, d_3 for which $\nabla_x f(x)^T d_1 > 0, \nabla_x f(x)^T d_2 < 0, \nabla_x f(x)^T d_3 = 0$? Can the same conclusions be drawn about the function of f?

Solution:

- d_1: Value of approximation goes up.
- d_2: Value of approximation goes down.
- d_3: Value of approximation stays the same.

The same can be said for f, but only in the immediate vicinity of the point x.

Intuition used here will be useful on the exam